Bioinformatics, 36(4), 2020, 1283–1285
doi: 10.1093/bioinformatics/btz716
Advance Access Publication Date: 3 October 2019
Applications Note

Sequence analysis

VikNGS: a C++ variant integration kit for next generation sequencing association analysis

Zeynep Baskurt [1,2,†], Scott Mastromatteo[1,2,†], Jiafen Gong[1,2], Richard F. Wintle[1,2], Stephen W. Scherer[1,2,3] and Lisa J. Strug[1,2,4,*]

[1]Program in Genetics and Genome Biology, Research Institute, The Hospital for Sick Children, Toronto, ON M5G0A4, Canada, [2]The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON M5G0A4, Canada, [3]McLaughlin Centre and Department of Molecular Genetics, University of Toronto, Toronto, ON M5G 0A4, Canada and [4]Division of Biostatistics and Department of Statistical Sciences, University of Toronto, Toronto, ON, M5T3M7, Canada

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Bonnie Berger

Received on January 28, 2019; revised on August 13, 2019; editorial decision on September 15, 2019; accepted on September 25, 2019

Abstract

Summary: Integration of next generation sequencing data (NGS) across different research studies can improve the power of genetic association testing by increasing sample size and can obviate the need for sequencing controls. If differential genotype uncertainty across studies is not accounted for, combining datasets can produce spurious association results. We developed the Variant Integration Kit for NGS (VikNGS), a fast cross-platform software package, to enable aggregation of several datasets for rare and common variant genetic association analysis of quantitative and binary traits with covariate adjustment. VikNGS also includes a graphical user interface, power simulation functionality and data visualization tools.
Availability and implementation: The VikNGS package can be downloaded at http://www.tcag.ca/tools/index.html.
Contact: lisa.strug@utoronto.ca
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genetic association studies contribute greatly to our understanding of complex traits. Association studies with whole genome sequencing (WGS) enables analysis of the full allele frequency spectrum and the decreasing cost of this technology continues to make it more accessible. Yet collaboration across study groups along with the integration of other publicly available WGS data is required to realize the statistical power necessary for the identification of associated loci.

Depending on resources and the scientific context, different groups may choose different experimental designs for their WGS studies. Using next generation sequencing (NGS) data, confidence in a variant call is dependent on, among other factors, the sequencing technology, read depth, error rate, base calling algorithm, alignment, SNP detection and genotype calling algorithms (Nielsen *et al.*, 2011). Naively pooling genotype calls from different WGS studies and performing an association test can result in spurious association findings due to the bias introduced by the differences in genotype call uncertainty across experimental designs. Previously, we showed the impact of differential read depth on case-control association studies when the cases and controls were sequenced with high and low read depth, respectively, and we developed a robust variance score statistic (RVS) to adjust for the resulting bias (Derkach *et al.*, 2014).

Here we introduce the Variant Integration Kit for NGS, VikNGS. VikNGS includes an extended version of the RVS framework (Derkach *et al.*, 2014) (vRVS; Supplementary Document) which enables integration of an arbitrary number of datasets that may have been sequenced using different experimental designs, and enables common and rare variant association testing for binary or quantitative traits with covariate adjustment. The software also provides rare and common variant association tests with genotype calls, e.g. CAST (Morgenthaler and Thilly, 2007) and SKAT (Wu *et al.*, 2011), as well as a simulation package for power and sample size estimation for study planning. The simulation package was used to conduct a comprehensive evaluation of the vRVS performance (Supplementary Document).

2 VikNGS package

VikNGS is a C++ software package that runs on Windows, Mac and Linux operating systems as a command line tool or using a graphical user interface. VikNGS is designed to run genetic association tests using NGS data and includes options for power analysis and interactive data visualizations. Association tests require a multi-sample VCF file and a tab-separated file containing individual-level

information as input (Supplementary Fig. S1). In practice the multi-sample VCF should be created by calling variants directly from the BAM sequence files. We did not investigate the impact of constructing the VCF by combining gVCFs but we expect similar results; varying the confidence threshold used can avoid preferential omission of rare variants if observed. The vRVS methodology can adjust for sequencing parameters, such as sequencing platform, read depth and coverage, but other systematic biases in the data (e.g. population stratification) can lead to spurious associations. A BED file can be supplied for rare variant analysis to enable the collapsing of variants within genes, exons or any arbitrary interval.

Given a set of filtering parameters and input files, VikNGS will parse, filter, collapse and perform association testing on variants concurrently (Supplementary Fig. S2). For each test performed, VikNGS will produce a *P*-value and output summarized variant information and results to a text file. If the graphical user interface is used, an interactive Manhattan-style plot will be produced (Fig. 1 and Supplementary Fig. S3) and variant-level information can be explored using a tabular view (Supplementary Figs S4 and S5). Data can also be simulated for power and sample size estimation (Supplementary Fig. S6). The results are visualized in Q-Q plots (Supplementary Fig. S7) or for power as a function of sample size (Supplementary Fig. S8).

2.1 Application to the genetics of cystic fibrosis (CF)

The causal *CFTR* locus. We applied VikNGS vRVS to an unbalanced case control study on chromosome 7 (273 241 variants with MAF > 5% analyzed) which contains the CF-causing cystic fibrosis transmembrane conductance regulator [*CFTR*; (Kerem *et al.*, 1989)]. We included NGS data from three independent studies: 101 individuals with CF sequenced by Complete Genomics at 30× (Panjwani *et al.*, 2018); 1927 non-CF individuals (TheUK10KConsortium, 2015) sequenced on Illumina HiSeq 2000 at 6.5×; and 379 individuals from the 1000 Genomes Project Phase 1 sequenced on a combination of ABI SOLiD and Illumina platforms at 4× (The Genomes Project, 2015). The analysis took 12 min to parse, filter and test the 99 GB VCF file (Supplementary Document; Supplementary Tables S28 and S29). Figure 1 shows a strong association at *CFTR*, as expected. We then investigated the distribution of *P*-values obtained from the short arm of chromosome 7, which should be distributed as uniform [0, 1] under the null hypothesis although long-range LD with *CFTR* will impact this slightly. The analysis with genotype calls has greater genomic inflation ($\lambda = 1.14$) than the vRVS ($\lambda = 1.06$), consistent with simulation results and the presumption that short arm variants are not associated with CF.

CF modifier gene *SLC26A9* in a non-CF population. Here we used VikNGS to implement an association test in a single study sequenced using one experimental design. Variants near *SLC26A9* were previously identified to contribute to lung function in older CF and non-CF populations (Strug *et al.*, 2018), and hypothesized to influence lung function by improving CFTR function. We used the ratio of the spirometry measures forced expiratory volume in 1 s and forced vital capacity as the quantitative lung function measure, which was available in 1927 non-CF participants in the UK10K data measured at the age of 8. Quantitative trait association analysis at the top CF-associated variant, rs4077468 (Sun *et al.*, 2012), also demonstrated evidence of association (vRVS $P = 0.0392$, and $P = 0.0375$ with genotype calls). This analysis offers further support for this locus contributing to lung function, even in 8-year-olds without CF.

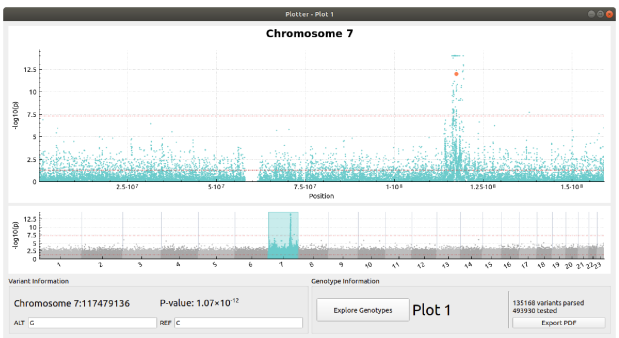

Fig. 1. Association test results from an unbalanced design with 101 individuals with cystic fibrosis and 2306 healthy controls. The red dotted lines indicate the $P = 0.05$ and $P = 5 \times 10^{-8}$ significance levels. The region around CFTR (chr7: 117 105 838–117 356 025) contains a cluster of genome-wide significant *P*-values. 1×10^{-14} is the minimum value calculated. Since our interest was in chromosome 7, variants on the other chromosomes were randomly generated under the null hypothesis of no association

3 Discussion

Traditional association testing with NGS uses called genotypes, the accuracy of which is highly dependent on read depth. Confounding read depth with case-control status will result in an inconsistent distribution of errors in genotype calls. For rare variants, the majority of errors will be mistaken heterozygous calls, resulting in what appears to be an enrichment of minor alleles in one group over another, leading to inflated significance. VikNGS enables rare and common variant association analyses on an arbitrary number of datasets for both binary and quantitative traits, allowing for covariate adjustment. Currently, the vRVS assumes that specified covariates are not strongly correlated with the genotype, G_{ij}. Simulation demonstrated that the vRVS controls Type I error and provides comparable power to analyses if the true genotypes were known, even for unbalanced case-control designs. In addition to vRVS, we implemented several common and rare variant association methods using genotype calls.

The vRVS methodology implemented in VikNGS will enable mega-analysis from large consortia despite differences in sequencing design. A current limitation of VikNGS is that it requires a multi-sample VCF file as input, which uses a variant calling step to identify polymorphic loci along the genome. (Hu *et al.*, 2016) note that rare variants can be indistinguishable from base calling errors in low read depth datasets which can lead to truly monomorphic sites being called as rare variants. Including these monomorphic sites in a rare variant test results in inflation in Type I error, even when using vRVS. As this is an issue with variant calling, one can filter out potentially monomorphic sites using an allele frequency filter or try using variant calling approaches such as PhredEM (Liao *et al.*, 2017).

VikNGS is a fast, user-friendly software package that can be run from the command line or using a visual interface on Windows, Linux or Mac operating systems. VikNGS is freely available under the MIT license at http://www.tcag.ca/tools/index.html with detailed documentation at https://vikngsdocs.readthedocs.io/en/latest/.

Funding

This work was supported by Genome Canada (OGI-138 to L.J.S. and S.W.S.) and the Natural Sciences and Engineering Council of Canada (2015-03742 to L.J.S.). This study makes use of data generated by the UK10K Consortium, derived from samples from ALSPAC study. A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310.

Conflict of Interest: none declared.

References

Derkach,A. *et al.* (2014) Association analysis using next-generation sequence data from publicly available control groups: the robust variance score statistic. *Bioinformatics*, **30**, 2179–2188.

Hu,Y.-J. *et al.* (2016) Testing rare variant association without calling genotypes allows for systematic differences in sequencing between cases and controls. *PLoS Genet.*, **12**, e1006040.

Kerem,B. *et al*. (1989) Identification of the cystic fibrosis gene: genetic analysis. *Science*, **245**, 1073–1080.

Liao,P. *et al*. (2017) PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. *Genet. Epidemiol*., **41**, 375–387.

Morgenthaler,S. and Thilly,W.G. (2007) A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res*., **615**, 28–56.

Nielsen,R. *et al*. (2011) Genotype and SNP calling from next-generation sequencing data. *Nat. Rev. Genet*., **12**, 443–451.

Panjwani,N. *et al*. (2018) Improving imputation in disease-relevant regions: lessons from cystic fibrosis. *NPJ Genomic Med*., **3**, 8.

Strug,L.J. *et al*. (2018) Recent advances in developing therapeutics for cystic fibrosis. *Hum. Mol. Genet*., **27**, R173–R186.

Sun,L. *et al*. (2012) Multiple apical plasma membrane constituents are associated with susceptibility to meconium ileus in individuals with cystic fibrosis. *Nat. Genet*., **44**, 562–569.

The Genomes Project. (2015) A global reference for human genetic variation. *Nature*, **526**, 68.

The UK10K Consortium. (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82.

Wu,M.C. *et al*. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet*., **89**, 82–93.